Overview of use Administrative Data in STS

Andrius Čiginas[1] and Daliutė Kavaliauskienė[2]

[1] Statistics Lithuania, Lithuania
e-mail: andrius.ciginas@stat.gov.lt

[2] Statistics Lithuania, Lithuania
e-mail: daliute.kavaliauskiene@stat.gov.lt

Abstract

In this paper we present the experience of Statistics Lithuania in the use of Administrative data in STS. The methodology proposed is based on analysis of data of four STS surveys.

1 Introduction

The political interest in administrative burden in general and statistical response burden in particular has never been higher than at the present time. Statistics is one of the priority areas in the ambitious goal set by the Commission of reducing the administrative burden on businesses by 25% within five years. The biggest reserve for the reduction of the burden on respondents is definitely the use of administrative data sources. On the other hand, a daily interest of survey statistician is the accuracy of statistical estimates and the administrative data can be considered as an auxiliary information, which may help to improve the quality of estimates. The Law on Statistics of the Republic of Lithuania gives statisticians the right to access administrative data for statistical purposes. Currently, Statistics Lithuania uses 110 administrative sources from different institutions and organizations. Most of the data received are aggregated (71 from 110). Primary data are taken directly from the State Tax Inspectorate (hereinafter STI), State Social Insurance Fund Board (SSIF) and some other institutions data bases. For the tax administration and for statistical purposes a data warehouse was established which is an integrated database on paid and declared taxes, financial-economic and other tax related indicators of tax payers (Joint Order signed by five state institutions). STI data warehous was created in the context of co-operation between the Central Customs and Tax Administration of the Danish Ministry of Taxation. However the possibility to use the administrative data sources does not solve the problem itself. The administrative databases are not so easy to use due to differences of indicators definitions, different deadlines for reporting and even technical problems.

1.1 STS surveys

The study is focused on 4 main STS surveys of enterprises.
The 2 of them are monthly: (i) domestic trade survey; (ii) industrial survey.
The other 2 are quarterly: (iii) service survey; (iv) construction survey.
The populations of those surveys include about 3/4 of active enterprises of Lithuania and in practice each enterprise belongs only to one of these surveys populations. For a

particular survey denote the population of interest by $\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$, where N is the size of survey population and u_i is the identifier of the population unit - enterprise.

Features which are common for all surveys:
1. The basic variable of interest for each survey is *income* of enterprises (i.e. monthly or quarterly income). Obviously, every separate survey aims to estimate *income* from different economical activity. The activity is typically dominant for a particular enterprise. For instance, we expect that the *income* of an enterprise involved into the construction survey, is basically from construction works done. But in reality many businesses can have income also from other economical activities. The basic indicator we need to estimate is the sum of *income*.
2. The sums of *income* have to be estimated not only for the whole survey population but also for particular levels of detail by economical activities (NACE) and size groups of enterprises. Typically size groups of enterprises are defined by the number of employees. The level of detail depends on national and (or) European needs. Thus by such classification we shall write $\mathcal{U} = \mathcal{U}_1 \cup \mathcal{U}_2 \cup \ldots \cup \mathcal{U}_G$, so that $\mathcal{U}_i \cap \mathcal{U}_j = \emptyset$, $i \neq j$, where $\mathcal{U}_j$ is the smallest group of enterprises or smallest estimation domain (SED) where the survey indicators need to be estimated.
3. The sampling design is usually prepared at the end of every year, when the estimation groups for the next year surveys are defined. Traditionally (and very naturally) the population of every survey is stratified accordingly to SED, i.e. the strata coincide with SDE. In separate case, when there are no planning requirements for grouping of enterprises by the number of employees, the SDE is actually one or another way stratified by the number of employees. Next the simple random sample is drawn independently from each stratum. The sample selected will be used for the whole next year.

For simplicity and without loss of generality we shall consider further single SDE $\mathcal{U}_j$ which possibly consists of some strata of sample design, i.e. $\mathcal{U}_j = U_1 \cup U_2 \cup \ldots \cup U_{S_j}$, where S_j is the number of strata in $\mathcal{U}_j$. This is because the estimates of indicators for combinations of SED is simply the sum of estimates in SED, which belongs to particular combination of SDE.

1.2 Commonly used auxiliary information

In order to determine the size of an enterprise at the sample selection stage and at the estimation stage the information on *annual income* of enterprises and *annual averaged number of employees* is used. These indicators are available at Statistical Business Register (SBR) of Statistics Lithuania.

Annual income - a mixture of information from different data sources:
- from STS surveys (sum of monthly or quarterly income),
- from annual statistical surveys,
- from STI data base, including VAT declarations data (the sum of monthly VAT data),
- other sources.

In the case when the data on income of the last year is not available, SBR uses the income data of enterprises, which may be few years old. It is important to note that the contribution of the VAT data is the largest, despite the fact that the priority of this data source is the last, because Statistics Lithuania can not directly control the quality of these data. Clearly, *annual income* is very convenient and useful variable for daily

work with STS surveys, because it allows to identify sizes of enterprises and therefore can be used for model-based estimation of sums of STS income, and sometimes can be used for the creation of sample designs of STS surveys. For instance, monthly industrial survey uses *annual income* at the estimation stage as auxiliary information variable for regression type estimator. For the remaining 3 STS surveys such application of *annual income* is more complicated, therefore traditionally H-T type estimator (which does not use auxiliary information) is applied. The main drawbacks of *annual income* data set are: a) it contains the data with different definitions; b) it contains data which may be few years old; c) the data set is not complete anyway - *annual income* is known for about 80% of active enterprises.

Annual averaged number of employees together with classification of economical activities (NACE) serves firstly for construction of sample design, i.e. for determination of bounds of population strata and for allocation of the sample size. This variable is also a mixture of different information but mostly it contains the data from SSIF. Differently from *annual income* the data from the data base of SSIF is leading in the sense that we have almost complete data set for our STS surveys populations - about 91% of information. The remaining part of information is the last year data or historical data from statistical surveys of Statistics Lithuania and the data from other external sources - about 8% of information.

1.3 Administrative data

Keeping in mind the structure of *annual income* contained in SBR very important auxiliary data source for STS surveys is VAT declarations. STS surveys usually use one derivative variable, which shows income of enterprise from all economical actions which are taxable and income of enterprise from a few economical actions which are not taxable (i.e. the tax rate is 0%). We shall call this variable *turnover*. The *turnover* data is monthly. Not all enterprises are the VAT payers. The *turnover* of enterprise have to exceed some certain limit during the last 12 months. Thus, roughly speaking, small enterprises are not the VAT payers and therefore only for about 65% of active enterprises the *turnover* is available monthly. The monthly or quarterly (the sum of 3 months) *turnover* became available few years ago, therefore it was less analyzed and compared with the corresponding *income* from monthly or quarterly STS surveys. Clearly, at first sight the definitions of *turnover* and *income* are different because *turnover* includes an income from almost all economical actions, i.e. it should "allways" exceed or be equal to *income* of enterprise, because the later typically contains an income from specific for particular survey economical actions. Statistical analysis in a sense confirms such opinion, but the reality is more complicated due: (i) measurement and other errors of the both VAT declarations *turnover* and STS *income* data; (ii) enterprise can overpay to STI and later the difference will be returned and conversely.

Number of employees from SSIF data base was the quarterly variable until the 2010. Since 2010 the monthly data is available. There are some differences between definitions of this variable and the corresponding data from statistical surveys but these differences are not significant. Also, differently from income of enterprise a variability in time of number of employees is much smaller.

2 The comparisons

In recent years Statistics Lithuania began more intensive search of methods which may allow to apply the data of STI and SSIF for purposes of STS surveys more efficiently. From the point of view of the definitions of *annual income* and *turnover* it can be expected that the last observed monthly or quarterly (the sum of 3 months) *turnover* should contain more "fresh" and homogenous information about income sizes of enterprises, although there is less data *turnover* available. That is, if our purpose is *income* of month or quarter t, then we would like to use the *turnover* of period t or period $t-1$.

Denote variables *income* of period t, *annual income* and (monthly or quarterly) *turnover* of period $t-1$ by y, x and z respectively. Then for enterprises of the particular STS survey population $\mathcal{U}$, the variable y (similarly x and z) attains values $y_1, y_2, \ldots, y_N$. We expect that y, x and z are linearly dependent. Therefore the first our (a bit rough) measure for comparisons is a coefficient of (Pearson) correlation. The analysis of 2005-2008 years data of STS surveys of interest showed that: 1) the estimate of correlation between y and z for most of the SED's exceeds the estimate of correlation between y and x; 2) the estimate of correlation between y and z for most of the SED's exceeds 0.9, i.e. in most of the cases we observe strong linear positive correlation.

Now consider an aggregated data of a particular year, i.e. variable y means annual *income* of enterprise from t-year STS survey (the sum of all months or quarters), variable x means $t-1$-year *annual income* and z means the sum of t-year months of *turnover*. Next define variables

$$u = \frac{z-y}{y} \qquad \text{and} \qquad v = \frac{x-y}{y}, \qquad \text{with} \qquad y \neq 0, \tag{1}$$

The purpose is to compare empirical distributions of these variables. Here is the few facts about these distributions.

1) If we exclude from our analysis 5% smallest and 5% largest values of u and v, then the averages of u and v for almost all SDE statisticaly significant do not differ from 0. Thus in a sense z and x do not differ from y.

2) Denote $q_1(u)$, $q_3(u)$ and $q_1(v)$, $q_3(v)$ the first and third empirical quartiles of distributions of u and v in particular SDE respectively. Then the interquartile range $IQR(v) = q_3(v) - q_1(v)$ for the most of SDE several times wider compared to the range $IQR(u) = q_3(u) - q_1(u)$. Thus x contains much more variability (with respect to y) compared to z.

3) For almost all SED the right "tail" of empirical distribution of u is more "heavy" compared to the left "tail". This is that we could expect by deffinitions of y and z.

Thus, there are some preliminary evidences about possible usefulness of *turnover*.

Similarly, few years ago in [3] the variable *number of employees* with number of employees from STS surveys were compared. The analysis showed that the data from SSIF are very close to the data from statistical surveys, for instance the coefficients of linear correlation for almost all SED are very close to 1. Also, as mentioned, the data from SSIF is available for very large parts of STS surveys populations. Therefore it was decided to exclude the variable number of employees from questioners of STS surveys. Thus our present knowledge about number of employees of enterprises is almost totally based on administrative data, i.e. based on SSIF data.

3 On editing of administrative data

In order to apply *turnover* data for estimation of sums of *income* we need more knowledge about relations between variables *turnover* and *income*. As we already know, these variables are linear dependent. The dependence can be modeled by simple regression line. The quality of model-based estimators, i.e. regression (or ratio) estimators of sums of *income* which are planned to be introduced in practice, is closely related to a quality of simple regression model. As mentioned above, we shall focus on single SED $\mathcal{D}$ of the particular STS survey. Assume that for sample units of $\mathcal{D} = \{u_1, u_2, \ldots, u_M\}$, where M is the size of domain, the *income* y_t values for period t are available. Denote them by $y_{t;1}, y_{t;2}, \ldots, y_{t;n}$, where n is the sample size in $\mathcal{D}$. Assuming that *turnover* data for period t is available, introduce variables $z_{t-k}, z_{t-k+1}, \ldots, z_t$, which mean a historical data of *turnover* from period $t-k$ to period t, where k is chosen number. (Note that if for particular STS survey the *turnover* data for period t are not available, we can consider variables $z_{t-k}, z_{t-k+1}, \ldots, z_{t-1}$.) Let $z_{t-j;1}, z_{t-j;2}, \ldots, z_{t-j;M}$ for $j = 0, 1, \ldots, k$ be the values of z_{t-j} for all units of $\mathcal{D}$. Clearly, values of z_t typically only for the part of enterprises of SED are available (see definition of *turnover*). Therefore we split up the SED $\mathcal{D}$ into two sub-domains $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$, where $\mathcal{D}_1 \cap \mathcal{D}_2 = \emptyset$, such that for all enterprises in $\mathcal{D}_1$ the values of z_t are known and for all enterprises in $\mathcal{D}_2$ the values of z_t are not known. Obviously, then the part of sample units will belong to $\mathcal{D}_1$ and the second part will belong to $\mathcal{D}_2$.

Now consider the data pairs $(y_{t;i_l}, z_{t;i_l})$, $l = 1, 2, \ldots, n_1$ of enterprises from sub-domain $\mathcal{D}_1$, where both of the components are observed. Here n_1 is the number of such pairs. Then consider simple linear regression model based on such pairs of observations

$$y_{t;i_l} = \tilde{\alpha} + \tilde{\beta} z_{t;i_l} + \tilde{\varepsilon}_{i_l}, \qquad l = 1, 2, \ldots, n_1, \tag{2}$$

where $\tilde{\varepsilon}_{i_l}$ is an error term. For the most of SED z_t explains y_t well, i.e. characteristics of (2) are nice. Also the interpretation of $\tilde{\alpha}$ and $\tilde{\beta}$ confirms the differences between the definitions of *income* and *turnover*: typically $\tilde{\alpha}$ statistically not significant differs from 0 and $\tilde{\beta} < 1$.

Our interest is the data of enterprises which (individualy) strongly affect the parameters $\tilde{\alpha}$ and $\tilde{\beta}$ of the model (2). Assume firstly that the data of *income* is "true", i.e. they are already edited. This assumption is quite natural because the survey data is collected for statistical purposes undergoing usual procedures of statistical editing including the secondary contacting with enterprises. Unfortunately, it is not the case with *turnover*. But there is the experience that strongly outlying data of *turnover* may be explained as follows: (i) The *turnover* of enterprise strongly exceeds the *income* because the enterprise sells the long-term tangible property. This fact can be explained sometimes by one of components of *turnover*, which shows income from not taxable actions (see definition of *turnover* in Subsection 1.3). (ii) Rough measurement errors or errors of data entry. (iii) The significant overpayment or underpayment of taxes.

Our method (or its modifications) below should detect outliers of such types.

Step1. Define new variable similarly as in (1) by $u_t = (z_t - y_t)/y_t$, where $y_t \neq 0$. Our interest is empirical distribution of this variable, which is based on observations $(y_{t;i_l}, z_{t;i_l})$, $l = 1, 2, \ldots, n_1$. Let $q_1(u_t)$, $q_3(u_t)$ be the first and third empirical quartiles of distribution of u_t. We shall assume that the value $u_{t;i_l}$ of u_t contains the potential

outlier (or error) if $u_{t;i_l} < q_1(u_t) - 3(q_3(u_t) - q_1(u_t))$ or $u_{t;i_l} > q_3(u_t) + 3(q_3(u_t) - q_1(u_t))$. Consider next for enterprise i_l, which has the potential outlying value of *turnover*, the historical data of *turnover* $z_{t-k;i_l}, z_{t-k+1;i_l}, \ldots, z_{t;i_l}$. Consider the empirical distribution based on these data. We will edit the value $z_{t;i_l}$ which corresponds $u_{t;i_l}$ if the absolute value of its **z**-value $|\mathbf{z}_{t;i_l}| = |z_{t;i_l} - \mu_{i_l}|/\sigma_{i_l} > 2$, where $\mu_{i_l} = \sum_{j=0}^{k} z_{t-j;i_l}/(k+1)$ and $\sigma_{i_l}^2 = \sum_{j=0}^{k}(z_{t-j;i_l} - \mu_{i_l})^2/k$. For such values we simply impute μ_{i_l} instead of $z_{t;i_l}$. Here (if $z_{t;i_l}$ is very rough error) also $\tilde{\mu}_{i_l} = \sum_{j=1}^{k} z_{t-j;i_l}/k$ can be imputed.

In order to apply regression or ratio estimator for estimation of sum of *income* we need to know also the sum of *turnover* in $\mathcal{D}_1$ (see (3) below). Clearly, for enterprises $\mathcal{D}_1^{'} = \mathcal{D}_1 \backslash \{u_{i_1}, u_{i_2}, \ldots, u_{i_{n_1}}\} = \{u_{j_1}, u_{j_2}, \ldots, u_{j_{m_1}}\}$, were m_1 is the size of set $\mathcal{D}_1^{'}$, the *income* values are not known and therefore the editing of the corresponding *turnover* is more complicated. Because in the set $\mathcal{D}_1^{'}$ the particular values of *turnover* are not important, our approach is the following.

Step2. Take the primary (not edited) values of *turnover* $z_{t;i_1}, z_{t;i_2}, \ldots, z_{t;i_{n_1}}$. For every enterprise u_{j_p}, $p = 1, 2, \ldots, m_1$ from the set $\mathcal{D}_1^{'}$ we can find the "nearest neighbour" u_{i_r} in the set $\{u_{i_1}, u_{i_2}, \ldots, u_{i_{n_1}}\}$, such that the absolute value of $a_{j_p;i_r} = z_{t;j_p} - z_{t;i_r}$ is the smallest among all absolute values (distances) $\{|a_{j_p;i_1}|, |a_{j_p;i_2}|, \ldots, |a_{j_p;i_{n_1}}|\}$. If there are a few "neighbours" with the same smallest distance, we will take one of them arbitrarily. If was decided in Step 1, that the *turnover* $z_{t;i_r}$ of "nearest neighbour" u_{i_r} should be edited, then we replace $z_{t;j_p}$ by $\left(1 + \frac{a_{j_p;i_r}}{z_{t;i_r}}\right)\mu_{i_r}$.

Remark. In the case then SED consists of several strata of sample design, the weights of sample design should be incorporated into method of editing described.

4 Estimation of indicators

The possibilities of using of *turnover* as auxiliary information for model-based estimation of sum of *income* are a bit different for different STS surveys. The *turnover* of the period t is not available for monthly STS surveys. Therefore we shall use the *turnover* of the period $t-1$. Similarly, for the quarterly survey on construction only the data for first two months of quarter t are available. In this case we shall use the sum of these two months of *turnover* as auxiliary information, or we shall use the sum of *turnover* of the last three months observed. Only for the quarterly survey on services the *turnover* is available for almost all VAT payers for all three months of period t.

Further the notation of the previous Section will be used. Also we assume, for simplicity, that the particular SED consists of only one strata of underlying sample design. Our approach is the application of combination of regression and direct (H-T type) estimates, i.e. in domain $\mathcal{D}_1$, where the auxiliary information for all enterprises is available, we use regression estimator and in domain $\mathcal{D}_2$ we use direct estimator. Denote in addition the *income* values for sample enterprises in domain $\mathcal{D}_2$ by $y_{t;k_1}, y_{t;k_2}, \ldots, y_{t;k_{n_2}}$, where n_2 is the sample size in domain $\mathcal{D}_2$. Following the notion of the previous Section for brevity denote by $M_1 = n_1 + m_1$ and by $M_2 = M - M_1$ the numbers of enterprises in $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively and denote by $A = \{i_1, \ldots, i_{n_1}, j_1, \ldots, j_{m_1}\}$ the set of identifying numbers of all enterprises of $\mathcal{D}_1$. Then the estimates of *income* sum for domains

$\mathcal{D}_1$ and $\mathcal{D}_2$ are

$$\hat{t}_{REG} = \frac{M_1}{n_1}\sum_{l=1}^{n_1} y_{t;i_l} + \tilde{\beta}\left(\sum_{l\in A} z_{t;l} - \frac{M_1}{n_1}\sum_{l=1}^{n_1} z_{t;i_l}\right) \quad \text{and} \quad \hat{t}_{HT} = \frac{M_2}{n_2}\sum_{l=1}^{n_2} y_{t;k_l} \tag{3}$$

respectively, where $\tilde{\beta}$ is the same as in (2). (Note that here, for instance, for monthly STS surveys we replace z_t by z_{t-1}.) Then the estimate for the whole $\mathcal{D}$ is $\hat{t}_{\mathcal{D}_1\cup\mathcal{D}_2} = \hat{t}_{REG} + \hat{t}_{HT}$. For comparison the usual direct (H-T) estimate for the whole $\mathcal{D}$ is

$$\hat{t}_{\mathcal{D}} = \frac{M}{n}\sum_{i=1}^{n} y_{t;i}. \tag{4}$$

The most popular accuracy measure of an estimate $\hat{t}$ is the estimate of coefficient of variation $\widehat{cv}(\hat{t}) = \sqrt{\widehat{\mathbf{Var}}\hat{t}}/\hat{t}$. For general theory of estimation of sums we refer to [4]. For practical applications (applications of special macro-command of SAS) of estimates as in (3) and their combinations we refer to [2]. Thus, the efficiency of competing estimators can be easily assessed by comparing the corresponding coefficients of variation. For the most of SED $\widehat{cv}(\hat{t}_{\mathcal{D}_1\cup\mathcal{D}_2})$ is considerably smaller than $\widehat{cv}(\hat{t}_{\mathcal{D}})$, i.e. (3) is more efficient compared to (4).
The last thing to do is to outline the situations when for particular SED the combination of (3) can not be applied: (i) the sample sizes n_1 or n_2 are too small. For instance, $n_1 < 10$ is too small for model (2) or $n_2 < 2$ is too small for estimation in domain $\mathcal{D}_2$; (ii) quality characteristics of the model (2) seem to be unacceptable. In the case of presence of (i) or (ii) we shall use usual estimate (4).

5 Concluding remarks

We have described quite in detail the methodological lines of use of VAT declarations data for estimation of the most important STS indicator - sum of *income*. It is important to note, that the problems of the use of other administrative data can be similar, i.e. differences in definitions, incompleteness of information, etc.
Differently from the use of *number of employees* of SSIF, the methodology presented here for VAT data is quite new and practical applications of it for all four STS surveys started from 2010. Before 2010 VAT data was used only for compensation of non-response.
The simulation study, performed using data of the year 2005-2008 in [1], showed that the new combined estimate gives the substantial decrease of coefficient of variation which is approximately equivalent to the following decrease of sample sizes in STS surveys (dependent on the year): survey of service enterprises - $24-29\%$; survey of construction enterprises - $25-34\%$; survey of industrial enterprises - $23-28\%$; survey of domestic trade enterprises - $13-30\%$. Thus, after successful application of the new methodology at 2010, the first reduction of sample sizes will start possibly from 2011.
Also we note that the common use of *turnover* for compensation of non-response should be performed very carefully, because the simulation study showed that in that case the estimates of sums of *income* are positively biased, what can be expected keeping in mind differences of definitions of *income* and *turnover*. For instance, given $4-6\%$ STS

survey non-response rate, without qualified editing of *turnover* (in this case the editing method in Section 3 does not work), the bias of estimate of sum of *income* is $2-4\%$.

References

[1] Čiginas A., *Use of STI data for STS in order to improve the efficiency of estimates and reduce the response burden.*, Report. Statistics Lithuania (In Lithuanian) (2009).

[2] Houbers M., Fangel S., *A quick guide to CLAN in 17 practical examples.*, Report. Statistics Denmark (2004).

[3] Kavaliauskas M., *Possibilities to reduce the statistical response burden by using data from SSIF data base.*, Report. Statistics Lithuania (In Lithuanian) (2006).

[4] Särndal C. E., Swensson, B., Wretman, J., *Model Assisted Survey Sampling.*, Springer-Verlag, New York (1992).